

Rule Acquisition for Cognitive Agents by Using Estimation of Distribution Algorithms

Tokue Nishimura and Hisashi Handa
Graduate School of Natural Science and Technology, Okayama University
Okayama 700-8530, JAPAN
email: handa@sdsc.it.okayama-u.ac.jp

Abstract—Cognitive Agents must be able to decide their actions based on their recognized states. In general, learning mechanisms are equipped for such agents in order to realize intelligent behaviors. In this paper, we propose a new Estimation of Distribution Algorithms (EDAs) which can acquire effective rules for cognitive agents. Basic calculation procedure of the EDAs is that 1) select better individuals, 2) estimate probabilistic models, and 3) sample new individuals. In the proposed method, instead of the use of individuals, input-output records in episodes are directory used for estimating the probabilistic model by Conditional Random Fields. Therefore, estimated probabilistic model can be regarded as policy so that new input-output records are generated by the interaction between the policy and environments. Computer simulations on Probabilistic Transition Problems show the effectiveness of the proposed method.

I. INTRODUCTION

The constitution of cognitive agents is done by acquiring rules for tasks through the interaction of their environment. Hence, this framework is formulated as Reinforcement Learning problems in general[1]. Conventional discrete Reinforcement Learning Algorithms provide us some theoretical support such that algorithms can achieve agents with optimal policy under Markov Decision Process [2]. However, in practical, such theoretical background may not be effective since some problems are far from Markov Decision Process. In addition, conventional Reinforcement Learning Algorithms are a kind of local search algorithms, where reward and value information is locally propagated into neighbor states. Therefore, in the case of huge problems, such value propagation takes much time to learn the whole problem space or cannot learn well. In the case of game or autonomous robots navigation, evolving neural networks and evolutionary fuzzy systems have attracted much attention recently [3][4].

Estimation of Distribution Algorithms (EDAs) are a promising Evolutionary Computation method, and have been studied by many researcher for the last decade [5][6]. The distinguished point of the EDAs is the use of probabilistic models, instead of crossover operator and mutation operator in conventional Evolutionary Computation. The probabilistic models in EDAs play crucial role in their optimization procedure so that the performance of EDAs is dramatically improved. Probabilistic models in EDAs try to model effective genetic information, i.e., schemata, so that conventional EDAs are applied to optimization problems, such as Ising spin glass problem, Max-Sat, Function optimization, attribute selection

problems so on [5][7][8][9]. As far as authors know, however, there are few research regarding to the application of EDAs to Reinforcement Learning because it require the interaction with environment.

Conditional Random Fields (CRFs) proposed by Lafferty *et al.* have been applied to segmentation problems in text processing and bio-informatics [10][11]. They employ undirected graphical probabilistic model, called Markov Network, in similar to Hidden Markov model. However, it is able to estimate conditional probability distributions, instead of joint probability distribution as in HMM. By realizing the estimation of conditional probability distributions, CRFs outperform HMM in natural text-processing area.

In this paper, a novel Estimation of Distribution Algorithm with Conditional Random Fields for solving Reinforcement Learning Problems is proposed. One of the primal features of the proposed method is direct estimation of Reinforcement Learning Agent's policy by using Conditional Random Fields. Therefore, the probabilistic model used in the proposed method does not represent effective partial solutions as in conventional EDA. Another feature is that a kind of undirected graphical probabilistic model is used in the proposed method. Recently, Markov Network is often used in EDA community [12][13][14][15]. However, they are using Markov Network for estimating joint probability distribution as in HMM.

The major contributions of this paper are summarized as follows:

- Solution of Reinforcement Learning Problems by using Estimation of Distribution Algorithms, i.e., promising method in ECs.
- Evolving Conditional Random Fields. Conventional CRFs are a sort of supervised learners, where knowledge is extracted only from database. The proposed method is able to explore new rules by means of Evolutionary search.

Remaining paper organization is described as follows: Section II begins by introducing Reinforcement Learning Problems. Section III briefly summarizes Conditional Random Fields. General calculation procedure of Estimation of Distribution Algorithms briefly introduced in IV. EDA-CRF proposed in V. Computer simulations on Probabilistic Transition Problems were carried out in Section VI.

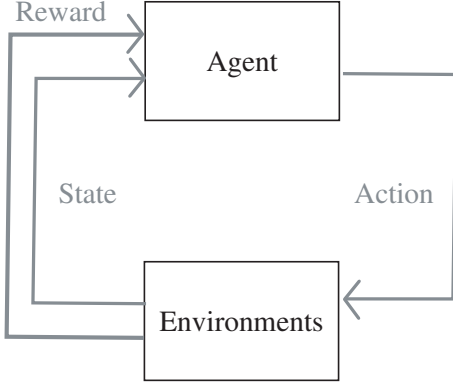


Fig. 1. Reinforcement Learning Problems

II. REINFORCEMENT LEARNING PROBLEMS

The reinforcement learning problem is the problem of learning from interaction with environments. Such interaction is composed of the perceptions of environments and the actions which affect both of agents and environments. Agents tries to maximize the total amount of reward which is given by environments as consequences of their actions. In other words, the task subject to agents is to acquire the policy $\pi(s, a)$ which yield a large amount of rewards, where s and a denote states recognized by agents, and actions taken by agents, respectively. The policy can be defined by using probabilistic formulation such as $P(a|s)$. Most Reinforcement Learning algorithms constitute value functions, e.g., state-action value function $Q(s, a)$ and state value function $V(s)$, instead of estimating the policy $\pi(s, a)$ directory. That is, in the case of conventional Reinforcement Learning algorithms, the policy $\pi(s, a)$ is approximated by value functions. In this paper, we employ Conditional Random Fields, introducing in the next section to estimate the policy $\pi(s, a)$.

III. CONDITIONAL RANDOM FIELDS

A. Overview

Conditional Random Fields (CRFs) were firstly proposed by Lafferty *et al.* in order to apply statistical learning into segmentation problems in text processing. The CRFs is a conditional distribution $P(\mathbf{y}|\mathbf{x})$ with an associated graphical structure. A distinguish point of the CRFs is to model such conditional distributions while Hidden Markov Models, which are traditionally used in broad area such as voice recognition, text processing and so on, estimate joint probability distributions $P(\mathbf{x}, \mathbf{y})$. This implies we do not have to consider the probabilistic distribution of inputs $P(\mathbf{x})$ since $P(\mathbf{x}, \mathbf{y}) = P(\mathbf{y}|\mathbf{x}) \cdot P(\mathbf{x})$. In general, $P(\mathbf{x})$ is unknown. Moreover, in the case of Reinforcement Learning Problems, it depends on agents' acts.

CRFs can be regarded as a mutant of Markov Network since CRFs use undirected graph model to represent probabilistic distribution. That is, variables in the problem are factorized in advance. Each clique (factor) is associated with a local

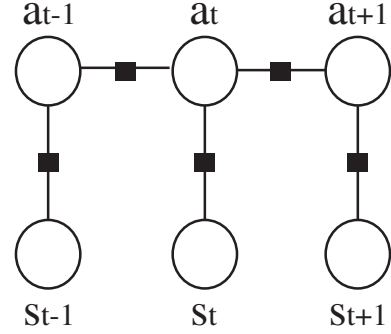


Fig. 2. Graphical model of Linear-Chain CRF

function. In the case of log-linear Markov Network, for instance, joint probability $P(\mathbf{y}, \mathbf{x})$ is represented as follows:

$$P(\mathbf{y}, \mathbf{x}) = \frac{1}{Z} \prod_A \Psi_A(\mathbf{y}_A, \mathbf{x}_A), \quad (1)$$

where Ψ_A denotes a local function for a set of variables $\mathbf{y}_A, \mathbf{x}_A \in \{\mathbf{y}, \mathbf{x}\}$. Z is a normalized factor which ensures that the distribution sums to 1:

$$Z = \sum_{\mathbf{y}, \mathbf{x}} \prod_A \Psi_A(\mathbf{y}_A, \mathbf{x}_A). \quad (2)$$

In the case of CRFs, following conditional probabilities $P(\mathbf{y}|\mathbf{x})$ are used:

$$\begin{aligned} P(\mathbf{y}|\mathbf{x}) &= \frac{P(\mathbf{y}, \mathbf{x})}{\sum_{\mathbf{y}'} P(\mathbf{y}', \mathbf{x})} = \frac{\frac{1}{Z} \prod_A \Psi_A(\mathbf{y}_A, \mathbf{x}_A)}{\sum_{\mathbf{y}'} \frac{1}{Z} \prod_A \Psi_A(\mathbf{y}'_A, \mathbf{x}_A)} \\ &= \frac{\prod_A \Psi_A(\mathbf{y}_A, \mathbf{x}_A)}{\sum_{\mathbf{y}'} \prod_A \Psi_A(\mathbf{y}'_A, \mathbf{x}_A)} \\ &= \frac{1}{Z(\mathbf{x})} \prod_A \Psi_A(\mathbf{y}_A, \mathbf{x}_A), \end{aligned} \quad (3)$$

where

$$Z(\mathbf{x}) = \sum_{\mathbf{y}'} \prod_A \Psi_A(\mathbf{y}'_A, \mathbf{x}_A).$$

B. Linear-chain CRF

The reason why we adopt CRFs to estimate probabilistic model is that CRFs can learn the whole sequence of input-output pairs, i.e., episode in the case of reinforcement learning. Although linear-chain CRF is the simplest form among CRFs, it still has such nature. Here, input and output variables \mathbf{x}, \mathbf{y} are substituted to a sequence of states $\mathbf{s} = \{s_1, s_2, \dots, s_t\}$ and a sequence of corresponding action $\mathbf{a} = \{a_1, a_2, \dots, a_t\}$, respectively. The linear-chain graphical model in this case is depicted in Fig. 2.

As we can see from this figure, the linear-chain CRFs factorize the set of variables \mathbf{s}, \mathbf{a} into state-action pairs (s_t, a_t) and transition of actions (a_{t-1}, a_t) . The local functions for each time step is defined as follows:

$$\Psi_k(\mathbf{a}_k, \mathbf{s}_k) = \exp(\lambda_k \cdot f_k(a_{t_{k-1}}, a_t, s_t)),$$

where $f_k(a_{t_{k-1}}, a_t, s_t)$ is a feature function which is either $\mathbf{1}_{\{a=a_{t-1}\}}\mathbf{1}_{\{a=a_t\}}$ or $\mathbf{1}_{\{a=a_t\}}\mathbf{1}_{\{s=s_t\}}$ and λ_k denotes parameter for factor k . At that time, equation (3) is rewritten as follows:

$$P(\mathbf{a}|\mathbf{s}) = \frac{1}{Z(\mathbf{s})} \exp \left\{ \sum_{k=1}^K \lambda_k f_k(a_{t-1}, a_t, s_t) \right\}. \quad (4)$$

C. Parameter Estimation

Suppose that N episodes are acquired to estimate the policy: $(\mathbf{s}^{(i)}, \mathbf{a}^{(i)})$, $(i = 1, \dots, N)$. Log likelihood method is used to estimate parameters λ_k :

$$\begin{aligned} l(\theta) &= \sum_{i=1}^N \log P(\mathbf{a}^{(i)}|\mathbf{s}^{(i)}) \\ &= \sum_{i=1}^N \sum_{t=1}^T \sum_{k=1}^K \lambda_k f_k(a_{t-1}, a_t, s_t) - \sum_{i=1}^N \log Z(\mathbf{s}^{(i)}) \end{aligned}$$

In order to calculate the optimal parameter θ , the partial derivative $\frac{\partial l}{\partial \lambda_k}$ of the above equation is used.

IV. ESTIMATION OF DISTRIBUTION ALGORITHMS

Estimation of Distribution Algorithms are a class of evolutionary algorithms which adopt probabilistic models to reproduce individuals in the next generation, instead of conventional crossover and mutation operations. The probabilistic model is represented by conditional probability distributions for each variable. This probabilistic model is estimated from the genetic information of selected individuals in the current generation. Fig. 3 shows the general process of EDAs. As depicted in this figure, the main calculation procedure of the EDAs is as follows:

- 1) Firstly, the N individuals are selected from the population in the previous generation.
- 2) Secondly, the probabilistic model is estimated from the genetic information of the selected individuals.
- 3) A new population whose size is M is then sampled by using the estimated probabilistic model.
- 4) Finally, the new population is evaluated.
- 5) Steps 1)-4) are iterated until stopping criterion is reached.

V. EDA-CRF FOR REINFORCEMENT LEARNING PROBLEMS

A. Overview

Fig. 4 depicts the calculation procedure of the proposed method, i.e., EDA-CRF. The procedure is summarized as follows:

- 1) Initial policy $\pi(s, a)$ is set to be of uniform distribution. That is, by according to the initial policy $\pi(s, a)$, agents moves randomly.
- 2) Agents interact with environments by using policy $\pi(s, a)$ until two episodes are generated. The episode denotes a sequence of pairs (state s_t , action a_t).
- 3) Better episode, i.e., the episode with much reward, among two episodes in the previous step is stored in

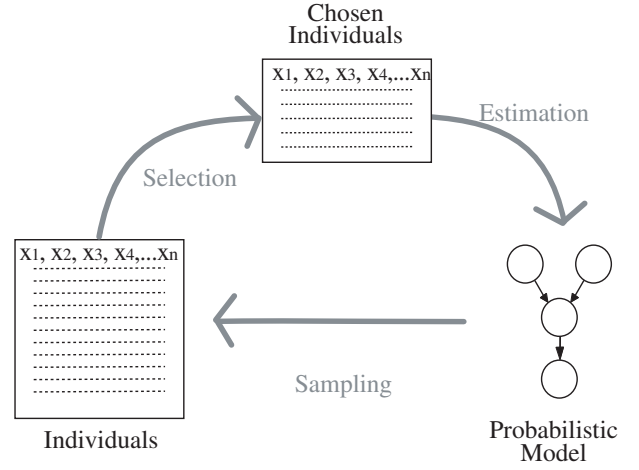


Fig. 3. Search Process by Estimation of Distribution Algorithms

episode database¹. Go back 2) until the number of chosen episodes reaches predefined constant value C_d .

- 4) A new policy $\pi(s, a)$ for the set of episodes in the database is estimated by CRF. After the estimation, all the episode data in the database is erased.
- 5) Go back to 2) until terminal conditions are met.

One of the main differences between conventional EDAs and EDA-CRF is the use of the estimated probabilistic model. Conventional EDAs employs the probabilistic model to generate individuals, i.e., solutions for given problems. Meanwhile, the probabilistic model estimated by CRF represents policy $\pi(s, a)$. In other words, the probabilistic model denotes solution itself.

Remarkable point of the proposed method is that in 1) in the above procedure, we are not intended to assume the uniform distributions for initial policy. If plenty of observation data, e.g. play-data by human, or episodes by conventional approach, is available, such data can be used to generate initial policy. This means that the proposed method can be easily incorporate human knowledge and can improve it by using evolutionary approach.

B. Interaction with Environments

As mentioned in the previous subsection, probabilistic model from selected episodes represents the policy of agents which decides actions for current situation. CRFs are originally used in text-processing and bio-informatics, where several output (y_1, y_2, \dots, y_i) should be decided against corresponding input (x_1, x_2, \dots, x_i) . Unfortunately, Reinforcement Learning Problems are not such a static problem. Every time step t , agents are subject to decide their outputs. Therefore, we employ factors regarding to deciding outputs are used to decide output. On the other hand, in the building probabilistic model phase, we take account into the whole sequence of pairs (states, actions). The factors used to choose an action a_t for

¹We employ "tournament selection" in GA [16]

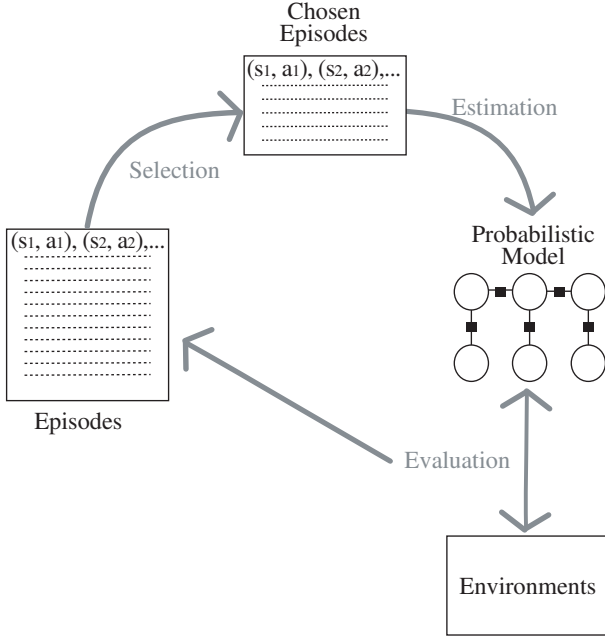


Fig. 4. Search Process by EDA-CRF

state s_t are $\mathbf{1}_{\{a=a_{t-1}\}}\mathbf{1}_{\{a=a_t\}}$ and $\mathbf{1}_{\{a=a_t\}}\mathbf{1}_{\{s=s_t\}}$. Hence, from equation (4)

$$P(a_t|s_t, a_{t-1}) = \frac{1}{Z(s_t)} \exp \left\{ \sum_{k=1}^K \lambda_k f_k(a_{t-1}, a_t, s_t) \right\}. \quad (5)$$

By using this probability and following equation, an action a at each time step t is chosen.

$$a = \operatorname{argmax} P(a_t|s_t, a_{t-1}).$$

Moreover, ϵ -greedy method is used in this paper so that, with probability ϵ , a new action is randomly chosen, instead of the above action. The parameter ϵ is set to be 0.05.

VI. EXPERIMENTS

A. Probabilistic Transition Problems

Probabilistic transition problem is introduced in order to investigate the effectiveness of the proposed method. A start point is located at center position. Agents can take two actions: left and right. By taking these actions, the state of agents is moved to one of adjacent states. However, with probability P_w , agents move to opposite direction of their chosen actions. There are two goals, one goal gives the agents reward $100/\text{count}$, where count denotes the number of steps until agents reaches to goals. Another one punishes the agents with negative reward $-100/\text{count}$. When agents reach to one of the goals, episode is finished. The number of intermediate states are examined 10 and 20.

B. Experimental Results

The proposed method, i.e., EDA-CRF is compared with Q-Learning, well-known Reinforcement Learning Algorithms. The parameters of Q-Learning are set that discount ratio

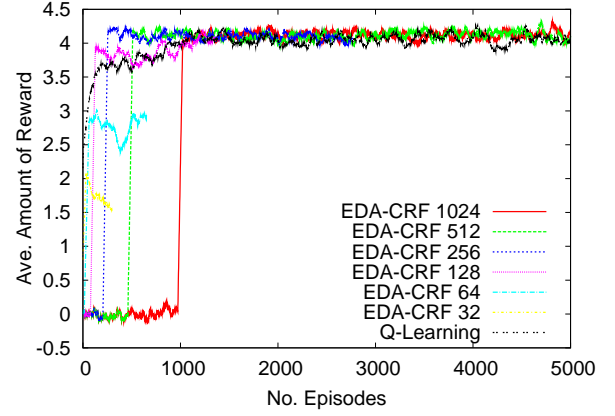


Fig. 8. Temporal changes of moving average of the amount of reward: $P_w = 0.3$

$\gamma = 0.9$, and learning ratio $\alpha = 0.1$. ϵ -greedy method is also adopted to Q-Learning.

Figs. 6 and 7 show the experimental results on the probabilistic transition problems with 10 and 20 intermediate states, respectively. In these figures, 4 graphs are plotted: experimental results on the error movement probability $P_w = 0.1, 0.2, 0.3$, and 0.4 . The number of episode per generation varies from 32 to 512. 10 generations are examined for all the EDA-CRFs. 10000 episode runs are carried out for Q-Learning. Except for $P_w = 0.3$ and 0.4 , and the number of episodes per generation is 32 or 64, the proposed method outperform the Q-Learning.

Fig. 8 show the temporal changes of the moving average of the amount of reward in the case of $P_w = 0.3$, the number of intermediate states is 10. X axis in this graph denotes the number of episodes. Drastic changes for EDA-CRFs are occurred at the end of the evaluation of initial individuals. Therefore, such drastic changes can be observed at corresponding population size. In other words, before such changes, agents employ random policy. The number of generation is set to be 10 for all the proposed methods so that the lines for EDA-CRFs 256, 128, 64, and 32, in the graph are terminated before the number of episodes is 5000. As you can see, most EDA-CRFs can effectively evolve the first generation. As we have seen in the previous figures, probability transition problems with $P_w = 0.3$ is difficult problem so that EDA-CRFs with the small number of episodes per generation, i.e., 32 and 64, are prematurely converged. Remarkable points in this figure is that EDA-CRF with 256 episodes per generation converges faster than Q-Learning. Moreover, as we can also see in Figure 6, EDA-CRFs with the large number of episodes per generation outperform Q-Learning.

VII. CONCLUSION

In this paper, Estimation of Distribution Algorithms with Conditional Random Fields (EDA-CRF) were proposed for solving Reinforcement Learning Problems. By using CRFs where the probabilistic model represents conditional probabilities, the proposed method is enable to cope with reinforce-

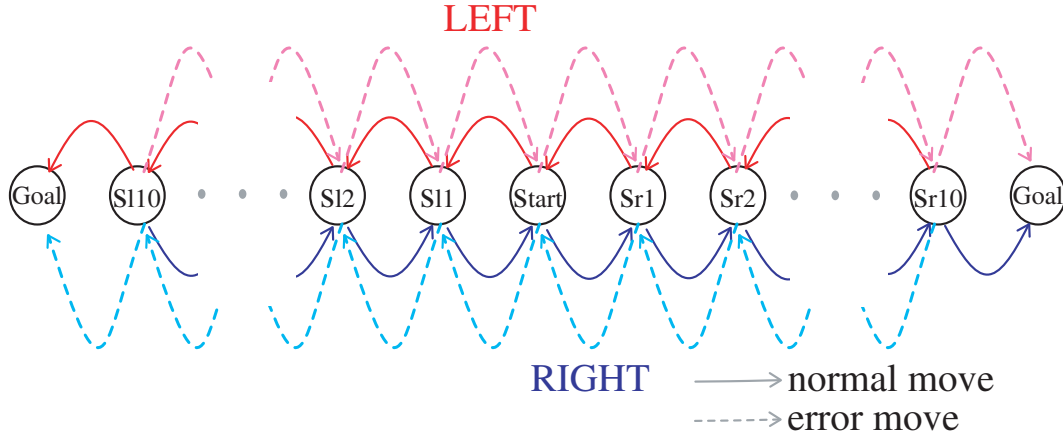


Fig. 5. A depiction of Probabilistic Transition Problems in the case that the number of intermediate state is 10

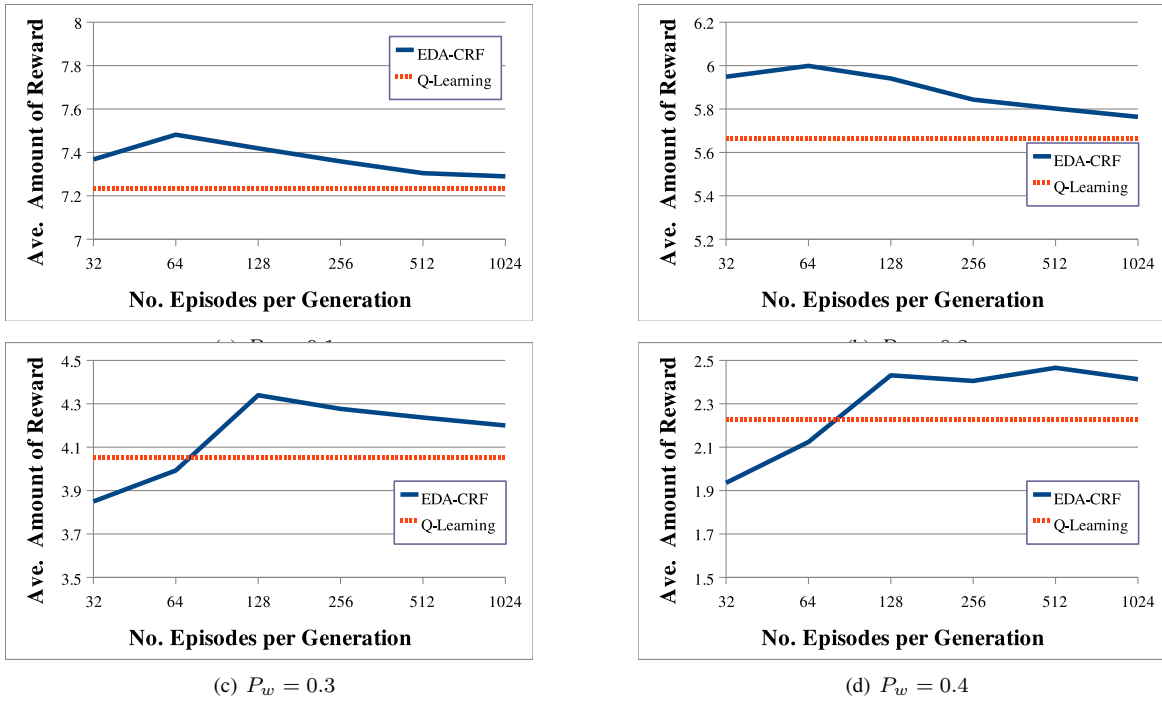


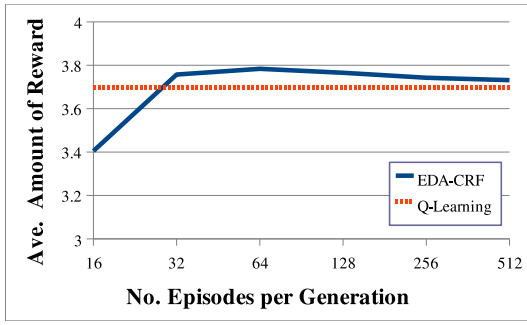
Fig. 6. Experimental results on probabilistic transition problems: The number of intermediate states is set to 10.

ment learning algorithms. Comparisons with Q-Learning on probabilistic transition problems show that EDA-CRF is the promising approach.

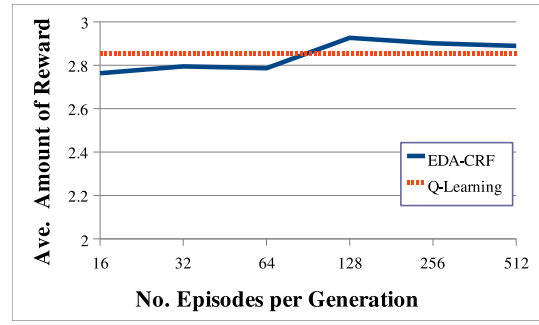
Finally, future works for the proposed method are addressed: Various kinds of problems, including continuous problems, should be examined to evaluate the effectiveness of EDA-CRF. Other factorizations in CRFs are needed to be investigated. We would like to capture the compatibility of problems and factorizations method in CRFs.

REFERENCES

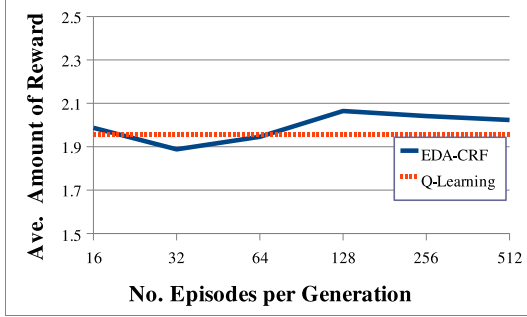
- [1] R. S. Sutton and A. G. Barto, *Reinforcement Learning; An Introduction*, The MIT Press, 1998.
- [2] C. J. C. H. Watkins, and P. Dayan. Technical note: Q-learning, *Machine Learning*, Vol. 8, pp. 279-292, 1992.
- [3] I. Ono, T. Nijo and N. Ono. A Genetic Algorithm for Automatically Designing Modular Reinforcement Learning Agents. *Proceedings of the 2000 Genetic and Evolutionary Computation Conference (GECCO2000)*, pp. 203-210, 2000.
- [4] H. Handa, and M. Isozaki. Evolutionary Fuzzy Systems for Generating Better Ms.PacMan Player. *Proceedings of the 2008 International Conference on Fuzzy Systems (Fuzz-IEEE 2008)*, pp. 2182-2185, 2008.
- [5] Larrañaga, P., et al. Combinatorial Optimization by Learning and Simulation of Bayesian, *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pp. 343-352, 2000.
- [6] H. Mühlenbein, T. Mahnig. FDA - a scalable evolutionary algorithms for the optimization of additively decomposed functions, *Evolutionary Computation*, Vol. 7, No. 4, pp. 353-376, 1999.
- [7] S. K. Shakya et al. Solving the ising spin glass problem using a bivariate eda based on markov random fields; *Proceedings of the 2006 IEEE Congress on Evolutionary Computation*, pp. 908-915, 2006.
- [8] I. Inza et al.: Filter versus wrapper gene selection approaches in DNA microarray domains; *Artificial Intelligence in Medicine*, Vol. 31, No. 2,



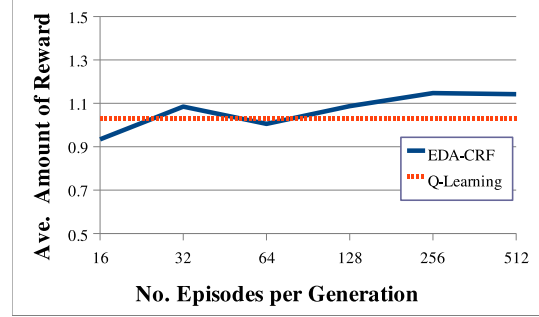
(a) $P_w = 0.1$



(b) $P_w = 0.2$



(c) $P_w = 0.3$



(d) $P_w = 0.4$

Fig. 7. Experimental results on probabilistic transition problems: The number of intermediate states is set to 20.

- pp. 91–103 (2004)
- [9] T. K. Paul, H. Iba: Gene Selection for Classification of Cancers using Probabilistic Model Building Genetic Algorithm; *BioSystems*, Vol. 82, No. 3, pp. 208–225 (2005)
 - [10] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proceedings of 18th International Conference on Machine Learning*, pp.282-289, 2001.
 - [11] C. Sutton, and A. McCallum. An Introduction to Conditional Random Fields for Relational Learning, In Lise Getoor and Ben Taskar, editors. *Introduction to Statistical Relational Learning*, MIT Press, 2007
 - [12] R. Santana. A Markov Network Based Factorized Distribution Algorithm for Optimization; *Proceedings of the 14th European Conference on Machine Learning* pp. 337-348, 2003.
 - [13] R. Santana. Estimation of distribution algorithms with Kikuchi approximations; *Evolutionary Computation*, Vol. 13, No. 1, pp. 67-97, 2005.
 - [14] S. K. Shakya, J. A. W. McCall, and D. F. Brown. Incorporating a metropolis method in a distribution estimation using markov random field algorithm; *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*, Vol. 3, pp. 2576-2583, 2005.
 - [15] S. K. Shakya *et al.* Using a Markov Network Model in a Univariate EDA: An Empirical Cost-Benefit Analysis; *Proceedings of the 2005 Genetic and Evolutionary Computation Conference*, pp. 727-734, 2005.
 - [16] D. E. Goldberg: *Genetic Algorithm in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.